ECOTERRA
Journal of Environmental Research and Protection

Estimation of TSS in the aeration tank of wastewater treatment plants

Bogdan Humoreanu, Ioan Naşcu, Ruben Crişan

Faculty of Automation and Computer Science, Technical University of Cluj-Napoca, Cluj-Napoca, Romania. Corresponding author: B. Humoreanu, Bogdan.Humoreanu@aut.utcluj.ro

Abstract. Industrial wastewater treatment covers the mechanisms and processes used to treat waters that have been contaminated by industrial or commercial activities prior to its release into the environment or its re-use. An estimation model for TSS in wastewater processing facility is presented. The model is built by data-mining algorithms based on WWTP data collected on a daily basis from SCADA system. The model performance by different algorithms is evaluated and the model built by the KStar algorithm has provided most accurate the estimations of TSS.
Key Words: TSS estimation, data-mining, data acquisition, WEKA, SCADA, estimation algorithms.

Introduction. Data mining (the analysis step of the "Knowledge Discovery and Data Mining" process, or KDD), an interdisciplinary subfield of computer science is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set using sophisticated mathematical algorithms to segment the data and evaluate the probability of future events (Peña-Ayala 2014; Dürrenmatt 2011; Finlay et al 2014; Dixon et al 2007; PhridviRaj & GuruRao 2014).

In wastewater treatment plants (WWTPs), much effort and money is invested in operating and maintaining dense plant-wide measuring networks. The network primarily serves as input for the advanced control scenarios that are implemented in the supervisory control and data acquisition system (SCADA) to satisfy the stringent effluent quality constraints. Due to new developments in information technology, long-term archiving has become practicable, and specialized process information systems are now available. The steadily growing amount of plant data available, however, is not systematically exploited for plant optimization because of the lack of specialized tools that allow operators and engineers alike to extract meaningful and valuable information efficiently from the massive amount of high-dimensional data. As a result, most information contained in the data is eventually lost (Airola et al 2011; Dürrenmatt 2011; Iverson & Randles 1987).

Examples of data-mining applications reported in the literature include the following: (a) prediction of the inlet and outlet biochemical oxygen demand (BOD) using multi-layered perceptions (MLPs), and function-linked, neural networks (FNNs); (b) modeling the impact of the biological treatment process with time-delay neural networks (TDNN); (c) predicting future values of influent flow rate using a k-step predictor; (d) estimation of flow patterns using auto-regressive with exogenous input (ARX) filters; (e) clustering based step-wise process estimation; and (f) rapid performance evaluation of WWTP using artificial neural network (Humoreanu 2013; Finlay et al 2014; Dixon et al 2007; PhridviRaj & GuruRao 2014; Guida et al 2007; Irfan et al 2013; Borra & Di Ciaccio 2010; Nenov 1995; Wei 2013).

Data acquisition and representation. Total suspended solids (TSS) give a measure of the turbidity of the water and are considered to be one of the major pollutants that contributes to the deterioration of water quality. As levels of TSS increase, a water body

begins to lose its ability to support a diversity of aquatic life. Suspended solids absorb heat from sunlight, which increases water temperature and subsequently decreases levels of dissolved oxygen (warmer water holds less oxygen than cooler water). Thus, it is imperative to know the values of influent TSS at future time horizons in order to maintain the desired characteristics of the effluent. The goal of this research project is estimation of TSS based on two aspects: (1) the input parameters used for the model development are: the pH of physico-chemical tank, the amount of dissolved oxygen in the aeration tank (DO_{air}), the amount of dissolved oxygen in the treated discharged water (DO_{ev}), the conductivity of the discharged water (σ), the pH of discharged water, the chemical oxygen demand of discharged water (COD); (2) model development for estimation of TSS in the aeration tank.

For each of the parameters listed above data were collected from SCADA system implemented for controlling and monitoring the processes in the treatment plant. Observed period is 01.06.2012 – 30.09.2012, sampling time is 2.66 hours and number of totally acquired data is 1046 for each parameter in order to estimate the value of TSS. To visualize the relationship between the values of TSS and parameters considered for TSS estimating scatter-plot diagrams are presented in Figures 1–6.

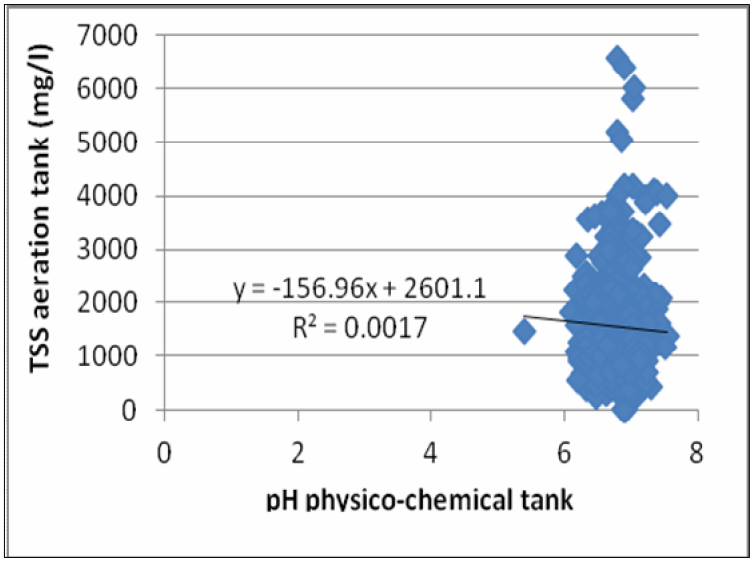

Figure 1. The relationship between TSS and pH from physico-chemical tank.

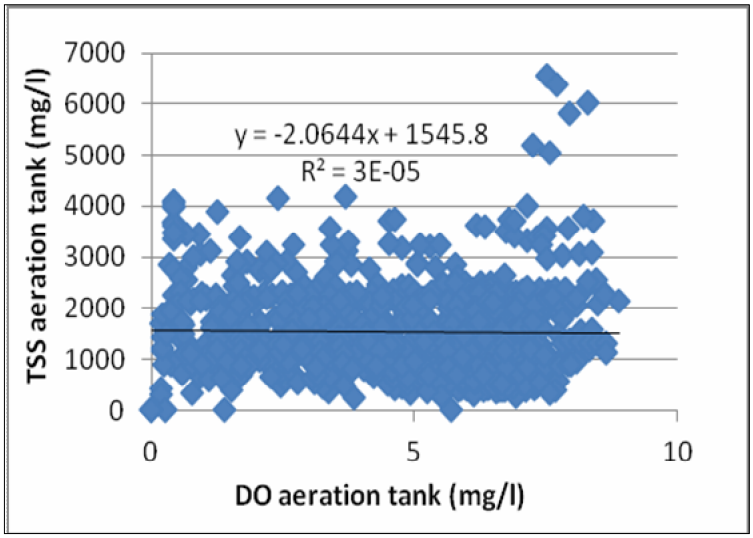

Figure 2. The relationship between TSS and DO from aeration tank.

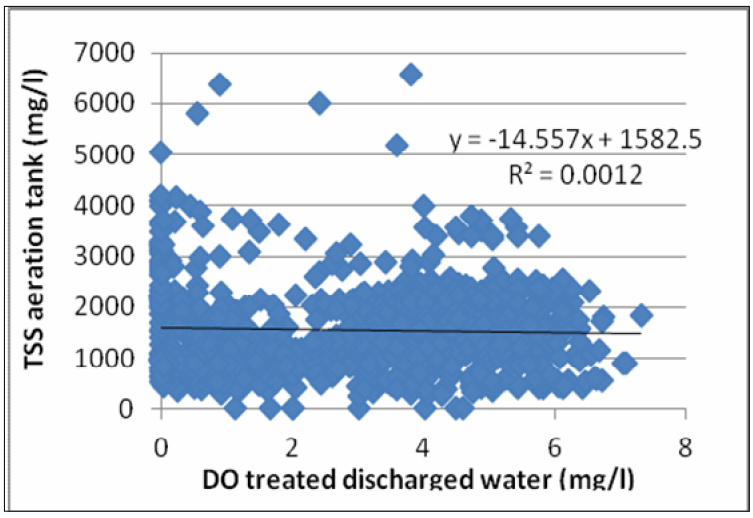

Figure 3. The relationship between TSS and DO from treated discharged water.

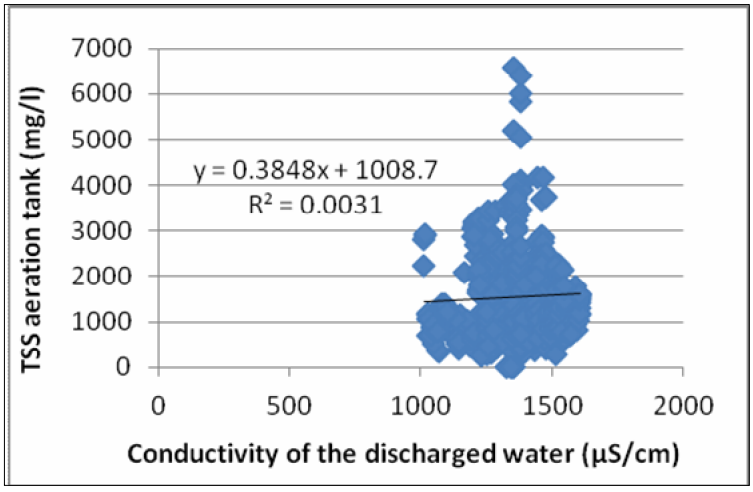

Figure 4. The relationship between TSS and conductivity of the discharged water.

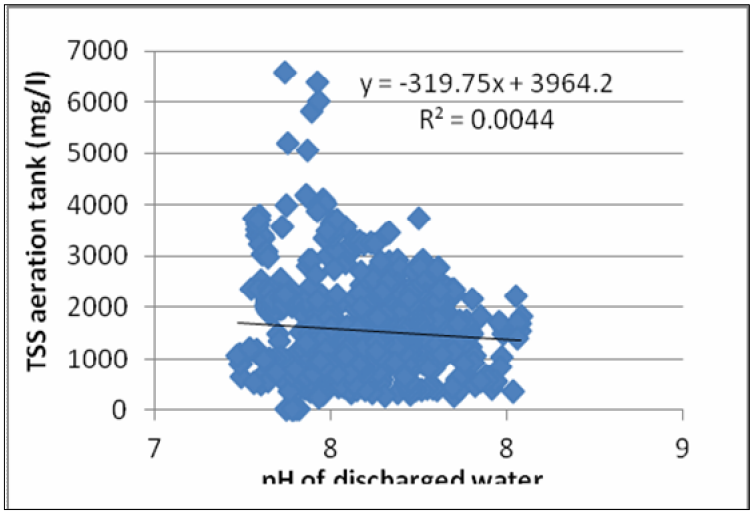

Figure 5. The relationship between TSS and pH of discharged water.

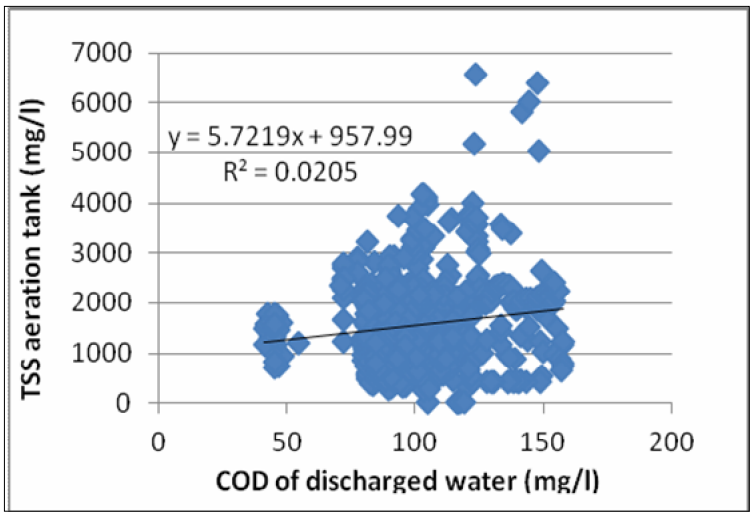

Figure 6. The relationship between TSS and COD of discharged water.

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.

The linear regression line has an equation of the form $Y=a+bX$, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when $x = 0$).

The strength of the linear association between the two variables is quantified by the correlation coefficient R. Given a set of observations (x_1,y_1), (x_2,y_2),...(x_n,y_n), the formula for computing the correlation coefficient is given by:

$$R = \frac{1}{n-1}\Sigma\left(\frac{x-\bar{x}}{S_x}\right)\left(\frac{y-\bar{y}}{S_y}\right)$$

where S_x and S_y are the sample standard deviations.

If the correlation coefficient is close to 1, it would indicates that the variables are positively linearly related and the scatter plot falls almost along a straight line with positive slope. For -1, it indicates that the variables are negatively linearly related and the scatter plot almost falls along a straight line with negative slope. And for zero, it would indicates a weak linear relationship between the variables (Peña-Ayala 2014; Kusiak et al 2013; Verma et al 2013; Airola et al 2011; Dürrenmatt 2011; Iverson & Randles 1987).

The square of the correlation coefficient, R^2 (R-squared), is a useful value in linear regression and is representing the coefficient of determination. R^2 indicates the proportionate amount of variation in the response variable y explained by the independent variable x in the linear regression model. The larger the R^2 is, the more variability is explained by the linear regression model (Haimi et al 2013; Humoreanu 2013).

R-squared is the proportion of the total sum of squares explained by the model and is a structure with two fields:

- Ordinary — ordinary (unadjusted) R-squared

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- Adjusted — R-squared adjusted for the number of coefficients

$$R_{adj}^2 = 1 - \frac{\left(\frac{n-1}{n-p}\right)SSE}{SST}$$

where SSE is the sum of squared error, SSR is the sum of squared regression, SST is the sum of squared total, n is the number of observations, and p is the number of regression coefficients (including the intercept). Because R-squared increases with added predictor

variables in the regression model, the adjusted R-squared adjusts for the number of predictor variables in the model. This makes it more useful for comparing models with a different number of predictors.

If R^2 is in range in between 0 and 1 indicates a perfect fit, and therefore, a very reliable model for future forecasts. If $R^2=0$ on the other hand, would indicate that the model fails to accurately model the dataset (http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm).

The low values of the coefficient of determination shown in the figures indicate a weak linear correlation between the input and output variables (parameters).

Thus, linear regression models are not suitable for estimation of TSS using the following parameters: the pH of physico-chemical tank, the amount of dissolved oxygen in the aeration tank (DO_{air}), the amount of dissolved oxygen in the treated discharged water (DO_{ev}), the conductivity of the discharged water (σ), the pH of discharged water and the chemical oxygen demand of discharged water (COD).

Selection algorithms for estimating TSS. For discovering association rules was used WEKA (Waikato Environment for Knowledge Analysis) platform which is a set of software for machine learning and data mining developed at the University of Waikato in New Zealand. Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. Given that the collected data have unbalanced distribution of classes (unbalanced data) that will affect the ability of correct classification algorithms to estimate TSS, an important aspect is the preparation of data for the classification process.

Classification algorithms are used to group multi-dimensional data into groups (clusters) defined algorithmically. This method is useful for quantifying large quantities of information, each group representing points having similar characteristics. In Weka platform were evaluated a series of algorithms to determine which algorithm is more suitable for the estimation of the TSS:

- M5P - implements base routines for generating M5 Model trees and rules;
- M5Rules - generates a decision list for regression problems using separate-and-conquer. In each iteration it builds a model tree using M5 and makes the "best" leaf into a rule;
- Decision Table - Class for building and using a simple decision table majority classifier;
- Bagging - Class for bagging a classifier to reduce variance and can do classification and regression depending on the base learner;
- RegressionbyDiscretization - a regression scheme that employs any classifier on a copy of the data that has the class attribute discretized. The predicted value is the expected value of the mean class value for each discretized interval (based on the predicted probabilities for each interval). This class now also supports conditional density estimation by building a univariate density estimator from the target values in the training data, weighted by the class probabilities;
- IBK - K-nearest neighbours classifier. Can select appropriate value of K based on cross-validation and can also do distance weighting;
- KStar - is an instance-based classifier, that is the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function. It differs from other instance-based learners in that it uses an entropy-based distance function.

Table 1 presents the results obtained from the use in WEKA platform of estimation algorithms listed above.

The training process is controlled by a cross-validation technique that is to randomly divide the original data set into three subsets: for training, for learning control (validation) and to assess the quality classifier (testing). Was used 10-fold cross-validation and the original sample is randomly partitioned into 10 equal size subsamples. Of the 10 subsamples, a single subsample is retained as the validation data for testing the model, and the remaining 9 subsamples are used as training data. The cross-validation process is then repeated 10 times (the folds), with each of the 10 subsamples

used exactly once as the validation data. The 10 results from the folds can then be averaged (or otherwise combined) to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once (Borra & Di Ciaccio 2010; Nenov 1995)

Table 1

Estimation of TSS using data-mining algorithms

Algorithm	*Testing Technique*	*The Mean Absolute Error (MAE)*	*Relative Absolute Error % (RAE)*	*Correlation coefficient*
M5P	Cross-Validation (*k = 10*)	177.85	33.23	0.84
M5Rules	Cross-Validation (*k = 10*)	191.16	35.72	0.81
Decision Table	Cross-Validation (*k = 10*)	185.67	34.69	0.84
Bagging	Cross-Validation (*k = 10*)	186.53	34.85	0.85
Regressionby Discretization	Cross-Validation (*k = 10*)	189.98	35.49	0.85
IBk	Cross-Validation (*k = 10*)	153.67	28.71	0.88
KStar	Cross-Validation (*k = 10*)	117.48	21.95	0.9

Based on the results in Table 1, the KStar algorithm outperforms the other algorithms by providing the lowest MAE and MRE errors. In the Figures 7-9 are represented the actual values and the estimated values of TSS in the aeration tank (Humoreanu 2013).

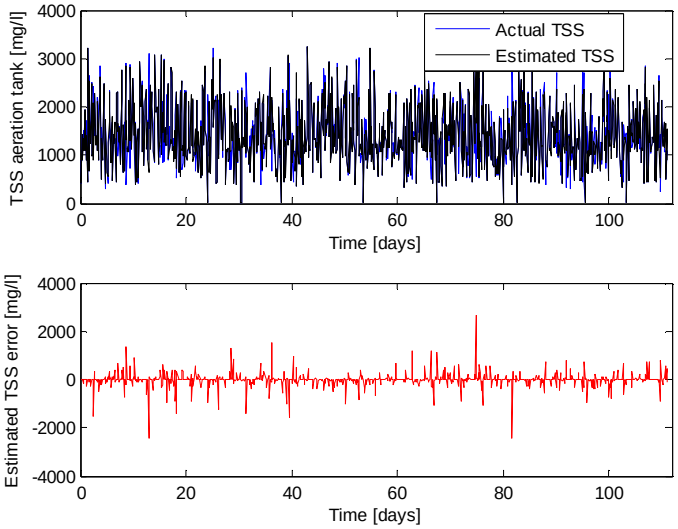

Figure 7. Actual and estimated values of TSS from the aeration tank in the period 01.06.2012 - 30.09.2012.

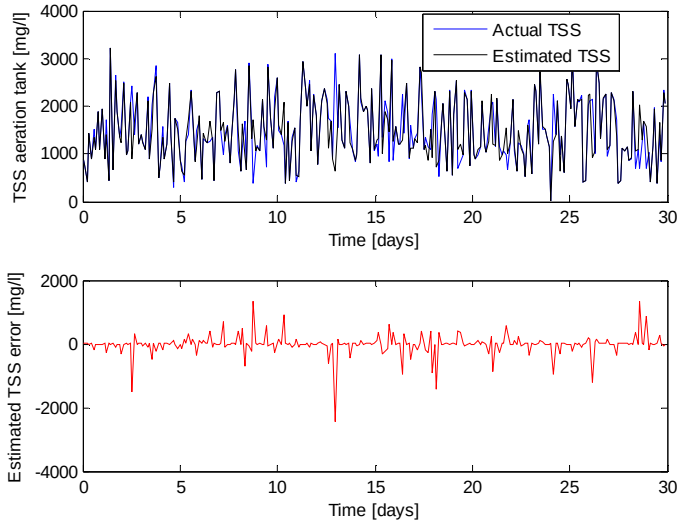

Figure 8. Actual and estimated values of TSS from the aeration tank in the period 01.06.2012 - 30.06.2012.

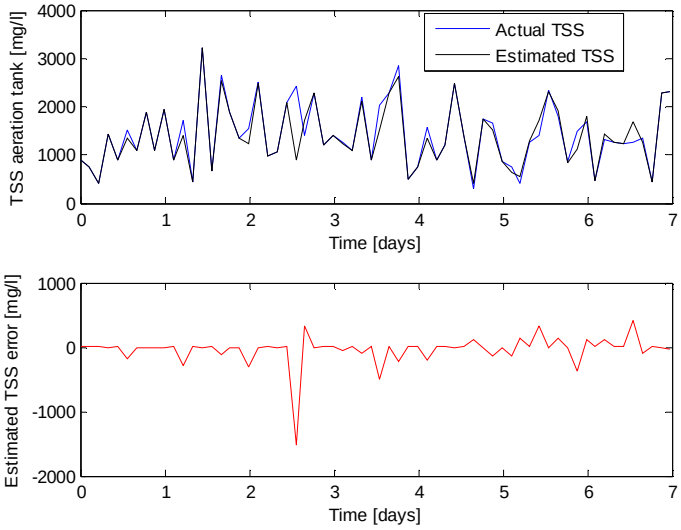

Figure 9. Actual and estimated values of TSS from the aeration tank in the period 01.06.2012 – 07.06.2012.

The error between actual and estimated values are shown in Figure 7 together a comparison between this two signals. The error range -2467 to +2673 it can be observed. Without the three picks the majority of error values varies in a short range.

Also, in Figures 8 and 9 are shown the actual and estimated TSS aeration tank and TSS estimated error. This figures represents a zoom in to initial time period (Figure 7). In Figure 8 the taken in to account period is one month (June 2012). The pick values, -2467 and 1343, can be observed. In Figure 9 are shown the comparison between the actual and estimated values of TSS together with the error signal. The error varies in range -1520 to 422.2, but the majority of values are concentrated around to zero. Just some picks can be observed.

The absolute estimated error of TSS from the aeration tank in period 01.06.2012 - 30.09.2012 is between 0 and 474.94 and is shown in Figure 10 (a). The variation error range are large, but the error signal is under 50 with some picks as it be observed from Figure 10 (a). The mean estimated error is better indicator than the range error and it is 10.59%. That mean error is considered acceptable for TSS estimation. Another indicator for the estimation procedure is the FIT parameter. The FIT is calculated using Normalized mean square error method with 2-norm of signals.

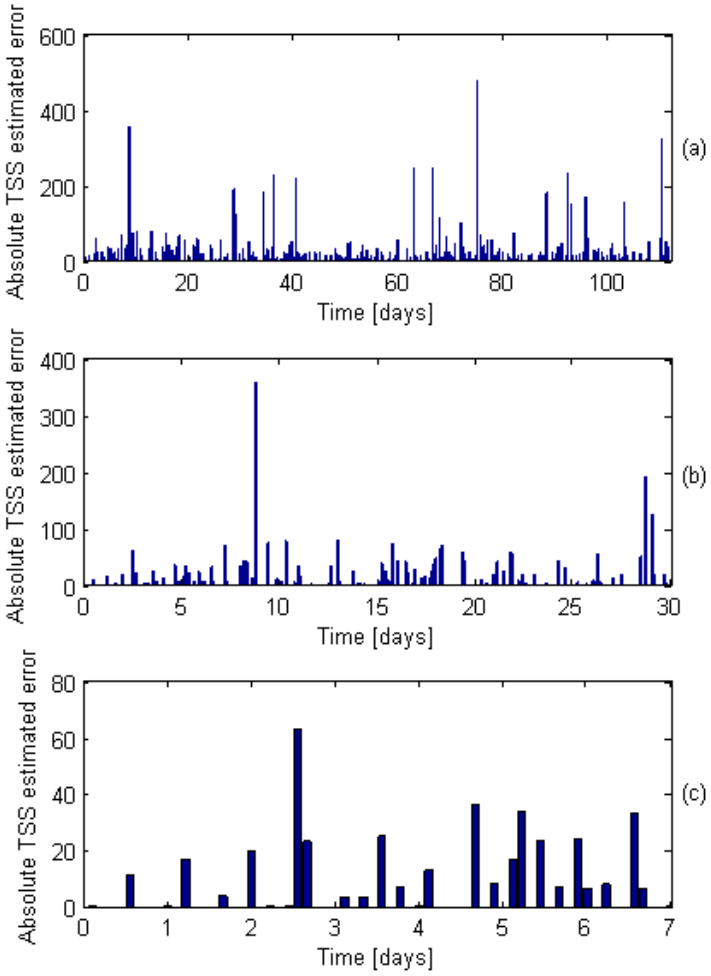

Figure 10. TSS estimated absolute error.

Thus,

$$FIT(i) = \left\| \frac{x(:,i) - xref(:,i)}{x(:,i) - mean(xref(:,i))} \right\|^2$$

where *x* is the estimated TSS signal and *xref* represents actual TSS.

The FIT calculated as forward is 80.74% that validate the estimation procedure.

For one month period (Figure 10 (b)) the FIT is 78.14%, a bit lees then FIT calculated on whole estimation period (01.06.2012 - 30.09.2012). The mean error is 11.28%, higher then the mean error achieved for whole estimation period.

Figure 10 (c) pointed out a comparison between the actual and estimated TSS together the estimated error signal for one week period. The mean error is 6.45% and the FIT is 87.52% that represents the best value of FIT among the ones calculated.

Conclusions. Plant disturbances on a wastewater treatment plant which discharges to a river can have disastrous consequences and, therefore, it is imperative that the monitoring system warn of significant changes as early as possible. An important benefit provided by SCADA system is the ease of analysis of plant data, which is continuously recorded. In this article has been introduced an approach based on data-mining techniques (extraction of knowledge from data) for analysis of water from the aeration tank. Values of the TSS sensor have been estimated based on the collected values from the SCADA system. This stage of the statistical analysis leads to the possibility of improving WWTP performance by allowing the construction of mathematical models that

characterize the water quality parameters and their implementation in the technological process.

Acknowledgements. This paper is supported by the project "Improvement of the doctoral studies quality in engineering science for development of the knowledge based society-QDOC" contract no. POSDRU/107/1.5/S/78534, project co-funded by the European Social Fund through the Sectorial Operational Program Human Resources 2007-2013.

References

Airola A., Pahikkala T., Waegeman W., De Baets B., Salakoski T., 2011 An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. Computational Statistics & Data Analysis 55(4):1828-1844.

Borra S., Di Ciaccio A., 2010 Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. Computational Statistics & Data Analysis 54(12):2976-2989.

Dixon M., Gallop J. R., Lambert S. C., Healy J. V., 2007 Experience with data mining for the anaerobic wastewater treatment process. Environmental Modelling & Software 22(3):315-322.

Dürrenmatt D. J., 2011 Data mining and data-driven modeling approaches to support wastewater treatment plant operation. ETH, an unpublished PhD thesis.

Finlay J., Pears R., Connor A. M., 2014 Data stream mining for predicting software build outcomes using source code metrics. Information and Software Technology 56(2):183-198.

Guida M., Mattei M., Della Rocca C., Melluso G., Meriç S., 2007 Optimization of alum-coagulation/flocculation for COD and TSS removal from five municipal wastewater. Desalination 211(1–3):113-127.

Haimi H., Mulas M., Corona F., Vahala R., 2013 Data-derived soft-sensors for biological wastewater treatment plants: an overview. Environmental Modelling & Software 47:88-107.

Humoreanu B., 2013 Scada systems for monitoring and control of wastewater treatmeant plants. Faculty of Automation and Computers, PhD Thesis.

Irfan M., Butt T., Imtiaz N., Abbas N., Khan R. A., Shafique A., 2013 The removal of COD, TSS and colour of black liquor by coagulation–flocculation process at optimized pH, settling and dosing rate. Arabian Journal of Chemistry. In Press, Corrected Proof, Available online 23 August 2013.

Iverson H. K., Randles R. H., 1987 Large sample properties of cross-validation assessment statistics. Journal of Statistical Planning and Inference 15:43-62.

Kusiak A., Zeng Y., Zhang Z., 2013 Modeling and analysis of pumps in a wastewater treatment plant: a data-mining approach. Engineering Applications of Artificial Intelligence 26(7):1643-1651.

Nenov V., 1995 TSS/BOD removal efficiency and cost comparison of chemical and biological wastewater treatment. Water Science and Technology 32(7):207-214.

Peña-Ayala A., 2014 Educational data mining: a survey and a data mining-based analysis of recent works. Expert Systems with Applications 41(4):1432-1462.

PhridviRaj M. S. B., GuruRao C. V., 2014 Data mining – past, present and future – a typical survey on data streams. Procedia Technology 12:255-263.

Verma A., Wei X., Kusiak A., 2013 Predicting the total suspended solids in wastewater: a data-mining approach. Engineering Applications of Artificial Intelligence 26(4):1366-1372.

Wei X., 2013 Modeling and optimization of wastewater treatment process with a data-driven approach. University of Iowa, PhD thesis, 135 pp.

*** http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm.

Received: 09 February 2014. Accepted: 27 February 2014. Published online: 31 March 2014.
Authors:
Bogdan Humoreanu, Faculty of Automation and Computer Science, Technical University of Cluj-Napoca, George Baritiu str., no. 26-28, 400027 Cluj-Napoca, Romania, e-mail: Bogdan.Humoreanu@aut.utcluj.ro
Ioan Naşcu, Faculty of Automation and Computer Science, Technical University of Cluj-Napoca, George Baritiu str., no. 26-28, 400027 Cluj-Napoca, Romania, e-mail: Ioan.Nascu@aut.utcluj.ro
Ruben Crişan, Faculty of Automation and Computer Science, Technical University of Cluj-Napoca, George Baritiu str., no. 26-28, 400027 Cluj-Napoca, Romania, e-mail: Ruben.Crisan@aut.cutcluj.ro

How to cite this article:
Humoreanu B., Naşcu I., Crişan R., 2014 Estimation of TSS in the aeration tank of wastewater treatment plants. Ecoterra 11(1):19-28.